IEEJ TRANSACTIONS ON ELECTRICAL AND ELECTRONIC ENGINEERING
*IEEJ Trans* 2021; **16**: 295–306
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI:10.1002/tee.23296

**Paper**

# ColAtt-Net: In Reducing the Ambiguity of Pedestrian Orientations on Attribute-Aware Semantic Segmentation Task

Mahmud Dwi Sulistiyo[*,**a], Non-member
Yasutomo Kawanishi[*], Non-member
Daisuke Deguchi[*], Non-member
Ichiro Ide[***,*], Non-member
Takatsugu Hirayama[****], Non-member
Hiroshi Murase[*], Non-member

Semantic segmentation has become one of the trending topics in the world of computer vision and deep learning. Recently, due to an increasing demand to solve a semantic segmentation task simultaneously with attribute recognition of objects, a new task named attribute-aware semantic segmentation has been introduced. Since the task requires to handle pixel-wise object class estimation with its attributes such as a pedestrian's body orientation, previous works had difficulties to handle ambiguous attributes such as body orientations in object-level, especially when segmenting the pedestrians with their attributes correctly. This paper proposes the ColAtt-Net that is an attribute-aware semantic segmentation model augmented by a column-wise mask branch to predict the pedestrians' orientations in the horizontal perspective of the input image. We firmly assume that the pedestrians captured by a car-mounted camera are distributed horizontally so that for each column of the input image, the pedestrian pixels can be labeled with one orientation uniformly. In the proposed method, we split the output of the base semantic segmentation model into two branches; one branch for segmenting the object categories, while the other one, as the novel column-wise attribute branch, is to map the recognition of pedestrian's orientations that are distributed horizontally. This method successfully enhances the performance of attribute-aware semantic segmentation by reducing the ambiguity on segmenting the pedestrian's orientation. Improvements on the pedestrian orientation segmentation are confidently shown by the proposed method in the experimental results, both in quantitative and qualitative views. This paper also discusses how the improved performance becomes an advantage in the autonomous driving system. 

**Keywords:** ColAtt-net; attribute-aware semantic segmentation; ambiguity of pedestrian orientations; column-wise prediction

*Received 11 June 2020; Revised 18 September 2020*

## 1. Introduction

Following scientific developments in the field of computer vision and deep learning, semantic segmentation has become a very hot topic discussed by researchers and practitioners. Simply put, semantic segmentation can be interpreted as a task to classify object categories and locate them in pixel-level of the captured image. Its applications are very broad, covering recent developments in the fields such as satellite imagery [1], medical imaging [2–4], robotics [5–7], and autonomous vehicle [8–12]. Various solutions and models have been competing with each other to become the best, such as FCN [13], SegNet [14,15] ICNet [16], Deeplab(s) [17–19], PSPNet [20], and many more. With the growing needs, those existing works become conventional as they only provide category names of the objects captured in the image. Additional information, such as object's attributes, which is simultaneously presented in the semantic segmentation outcome might give a better scene understanding. For that, recently, we have introduced the attribute-aware semantic segmentation task [21–23] to simultaneously collaborate semantic segmentation and attribute recognition tasks. The study focused on a pedestrian and its body orientations as the targeted object and its attributes. There are two types of class in this study; 'object' class which includes *road*, *building*, *car*, *person*, so on, and 'attribute' class which includes four pedestrian orientations, consisting of *back*, *right*, *front*, and *left*. Compared to the conventional semantic segmentation task, it improves not only the performance in general, but also enriches the output information.

However, for classifying the pedestrian orientations in pixel-level, the difficulty is to obtain a stable result in the orientation segmentation. The ambiguity of pedestrian attributes in object-level occurs when there are two or more body orientations encountered in one pedestrian; For example, the upper and

[a] Correspondence to: Mahmud Dwi Sulistiyo, E-mail: mahmuds@murase.is.i.nagoya-u.ac.jp

[*] Graduate School of Informatics, Nagoya University, Nagoya, Japan
[**] School of Computing, Telkom University, Bandung, Indonesia
[***] Mathematical and Data Science Center, Nagoya University, Nagoya, Japan
[****] Institutes of Innovation for Future Society, Nagoya University, Nagoya, Japan

lower body parts of a pedestrian may be segmented as different orientations. If we observe the images captured by a car-mounted camera as well as the reality in traffic situations, pedestrians are typically distributed horizontally. It is unlikely that two or more pedestrians are located vertically in spatial dimension. We can assume that predicting a pixel-wise orientation can cause ambiguity in some parts of the pedestrian's body. Whereas performing a column-wise prediction on pedestrian orientations will be more effective to avoid the ambiguity of orientation in one pedestrian body. With these assumptions, it is sufficient to predict one orientation class uniformly for pedestrian pixels along each column in the input image. This idea becomes a key basis of the method proposed in this paper to handle the ambiguity in segmenting the pedestrian orientations along with the object classes simultaneously. For some applications in an autonomous driving system, reducing the ambiguities of pedestrian orientations recognition is very important; Thus, it becomes our concern in this paper.

In this paper, the ColAtt-Net is introduced as a newly proposed model for the attribute-aware semantic segmentation task. The ColAtt-Net treats all pedestrian pixels in each column to belong to the same orientation class. This model divides the tasks into two parts; The first part is in charge of segmenting images into predetermined object categories, while the second part is aimed at predicting the orientation labels of the segmented pedestrian. The proposed method provides advantages over another past work [24)] that executes semantic segmentation, object detection, and attribute recognition in a series of separated processes; the ColAtt-Net is a multi-task learning-based model trained in an end-to-end process to perform these three stages simultaneously in a network. This means, the two branches of the model that perform the object categories segmentation and the pedestrian orientations prediction tasks are trained together with a combined loss function. In the end, the two outputs need to be post-processed to combine them.

We conducted some experiments to optimally train the ColAtt-Net and compare the results with the baseline method to show its advantages, both in quantitative and qualitative views. Figure 1 exemplifies that the ColAtt-Net produces a better segmentation as it successfully eliminates the ambiguity in segmenting the pedestrian orientations.

To sum up, our contributions in this paper are as follows:

- We propose the ColAtt-Net as a multi-task learning framework for the attribute-aware semantic segmentation task by introducing a column-wise pedestrian orientation prediction module to particularly improve the performance of an object's attributes segmentation.
- We provide experimental results to observe the optimal parameters for training the ColAtt-Net model.
- We show the improvements made by the proposed method compared to the baseline method both in quantitative metric and in qualitative perspective.

The rest of this paper explores existing works related to this study in Section 2 and details of the proposed methods in Section 3. In Section 4, we show the experimental results and discuss the advantages of the proposed method. Finally, Section 5 concludes the paper with possible future work.

## 2. Related Work

### 2.1. Attribute recognition: Pedestrian's orientation

In a traffic situation, many of the same traffic objects, such as pedestrians, are captured, but show various attributes and behaviors that may influence the development of applications [25–27] for Intelligent Transportation Systems (ITS). Some past studies [28–30] have conducted pedestrian attributes recognition, but were designed for images with single-cropped condition and captured by a surveillance camera. Among various types of existing attributes, 'body orientation' is an important attribute of a pedestrian for traffic scene understanding. For instance, for an autonomous driving system, estimating a pedestrian's body orientation is required before detecting the pedestrian's crossing intention [31], which is important to perform further actions. For this reason, this research focuses on optimizing the recognition of pedestrian orientations.

**2.2. Semantic segmentation** In computer vision, various techniques for interpreting an image have been developed. The variety starts from the most global context to the more detailed ones, including image classification, object detection, and semantic segmentation. A computer may obtain a comprehensive vision with the semantic segmentation as it classifies all categories of objects covered by the image in the smallest scale, that is, pixel-level.

There are numerous datasets publicly available for the semantic segmentation task, such as CamVid [32], KITTI [33], Mapillary Vistas [34], Cityscapes [35,36], and so on. From there, lots of methods and breakthroughs have born in recent years, as some of those mentioned in Section 1. Some review papers [37–39] summarize recent progress of the semantic segmentation. The solutions offered are mostly segmentation models based on Convolutional Neural Networks (CNN) and deep learning techniques since this task requires very detailed and complex classification capabilities.

**2.3. Attribute-aware semantic segmentation** With the development of some creative ideas, such as instance segmentation [40], multi-human parsing [41], and panoptic segmentation [42], conventional semantic segmentation has become inadequate to meet the needs of the system to be applied. Recently, attribute-aware semantic segmentation [21] [22] is introduced as a task that solves the increasing demand to simultaneously run the semantic segmentation and the object's attribute recognition. The CityWalks dataset [43] was constructed for this task. One application is in an

Input Image | Ground Truth | Previous Work[22] | Proposed Work

Fig. 1. Comparing outputs between the previous work [22] and the proposed work that proposes ColAtt-net

autonomous driving system which tries to provide a better traffic scene understanding for purposes related to the vehicle's motion planner and anticipation.

In the previous works [21–23], the segmentation model predicted body orientation as the pedestrian's attribute. However, it is a common problem in a semantic segmentation task to segment all pixels of the input image into correct classes, including the attribute values. The misclassification in pixel-level can cause certain parts of a pedestrian annotated with an incorrect orientation class or a different class from the other body parts. This makes the attribute-aware semantic segmentation task face ambiguity because two or more different orientations appear in one pedestrian instance. For example, the upper part of a person is segmented as a pedestrian walking to the left, while the lower part to the right. An existing study [44] has tried to address a similar issue but in an object detection task. This kind of problem, especially in attribute-aware semantic segmentation task, will be minimized in the current study.

### 2.4. Multi-task learning semantic segmentation

The problem of attribute-aware semantic segmentation can be seen as a multi-task learning (MTL) framework. Basically, there are two tasks; segmenting the object categories and segmenting the object's attributes, which are carried out simultaneously. However, there are still few existing studies using the MTL framework that address this particular problem. In our preliminary study [24], the attribute-aware semantic segmentation task was executed in three separated conventional stages; semantic segmentation, pedestrian detection, and attribute recognition. Another previous work [22] introduced an MTL-based model as a comparative method. However, the results were still not optimal; The ambiguity in pedestrian orientation segmentation occurred as the result of inaccuracies in the pixel-level classification. Therefore, this paper proposes a novel method using the MTL framework for attribute-aware semantic segmentation that can reduce the potential ambiguity in segmenting the pedestrian orientations.

## 3. Proposed Method: ColAtt-Net

This paper proposes the ColAtt-Net, an attribute-aware semantic segmentation model based on the MTL framework that splits the network's output into two branches; the object category segmentation and the column-wise prediction of pedestrian's attribute. Figure 2 shows the general stages in developing the proposed model. The two output branches are trained during the training stage to perform these two tasks simultaneously. Meanwhile, in the testing stage, the network has two steps; the inference step that yields two network's outputs including the object segmentation and the column-wise orientation prediction, and the post-processing step which collaborates the two inference outputs to obtain the final attribute-aware segmentation output. Before going through the main part of the proposed method, we introduce a preliminary investigation that becomes the most important cue to understand the novelty in this work.

### 3.1. Preliminary investigation

Figure 3 displays an output of the attribute-aware semantic segmentation by the method in the previous work [22]. From here, we can find pixels with

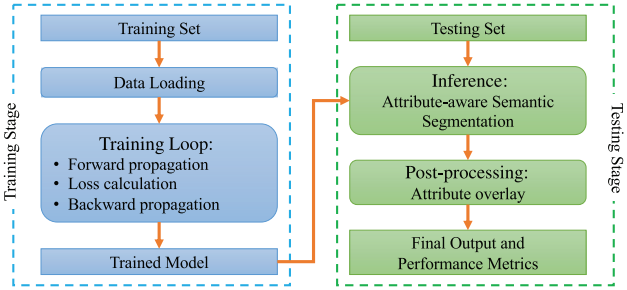


Fig. 2. General view in building the ColAtt-net model

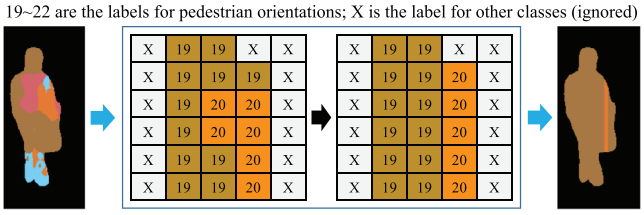


Fig. 3. Before and after applying a preliminary post-processing to an output produced in the previous work [22]

different orientations in some parts of a pedestrian. Although only a small part, it still causes ambiguity in recognizing the pedestrian's attribute in object-level. Suppose that the minor pixels with different orientations are incorrectly classified and overwritten with the major orientation label. In addition, since pedestrians in the input image are assumed to be distributed horizontally, one orientation attribute is sufficient for all pedestrian pixels in each column of the image. Thus, we can post-process the segmentation output by re-annotating the pedestrian pixels along each column with the most frequent orientation predicted in the corresponding column. This mechanism is illustrated in Fig. 3; The number of ambiguous pixels is significantly reduced.

That condition typically occurs in the outputs of attribute-aware semantic segmentation task, and thus, this post-processing mechanism is applicable for general cases. Table I shows that applying this simple method to three typical results from the previous work [22] is capable of increasing the performance metrics quantitatively. However, this effort may fail if the correct orientation does not sufficiently dominate the pedestrian's body part in the vertical spatial dimension. Adding this simple post-processing may even worsen the results; It depends on the output of pixel-wise prediction of pedestrian orientations. Therefore, a model based on the multi-task learning framework can be developed. The segmentation model needs to be specifically trained to have two simultaneous abilities: perform semantic segmentation for object categories and predict the orientation for each pedestrian as uniformly as possible.

Based on this preliminary investigation, we draw some conclusions as follows:

- Applying the idea that one orientation class is sufficient for one column of the image can be effective to reduce ambiguity in segmenting the pedestrian orientations.
- To avoid failures, the tasks of object segmentation and pedestrian orientation prediction can be performed simultaneously in a separate sub-task.

Table I. Segmentation performance with and without the preliminary post-processing mechanism

| # | Post-Pro. | $mIoU_4$ | $IoU_{back}$ | $IoU_{right}$ | $IoU_{front}$ | $IoU_{left}$ |
|---|---|---|---|---|---|---|
| 1 | No | 45.63 | 43.23 | 50.95 | 44.11 | 44.24 |
| | **Yes** | **46.71** | **44.70** | **52.60** | **44.59** | **44.96** |
| 2 | No | 48.71 | 45.91 | 54.17 | 44.55 | 50.222 |
| | **Yes** | **49.64** | **46.89** | **55.41** | **45.08** | **51.18** |
| 3 | No | 41.01 | 44.19 | 44.53 | 43.93 | 31.37 |
| | **Yes** | **42.48** | **45.26** | **46.19** | **44.19** | **34.29** |

- This encourages us to develop a multi-task learning-based model for attribute-aware semantic segmentation task which is trainable in an end-to-end process.

**3.2. ColAtt-Net's architecture** In this study, we propose the ColAtt-Net, which stands for 'Column-wise Attribute-aware semantic segmentation Network'. The ColAtt-Net model is built for the attribute-aware semantic segmentation task by introducing an augmenting column-wise orientation network branch to enhance the performance, especially in reducing the ambiguity of pedestrian orientation prediction. This method is proposed according to the following assumptions:

- Ambiguity in segmenting the pedestrian's orientation occurs when two or more orientations are predicted to one pedestrian instance, for example, the upper and the lower body parts of a pedestrian have different orientations.
- In the traffic scene images, pixels corresponding to each pedestrian captured by a car-mounted camera are assumed to be distributed horizontally.
- Two or more pedestrians can exist in the same vertical axis if they have different distances from the camera. In such a condition, the closer pedestrians will appear visually dominating, and in reality, they will be prioritized for collision prevention.
- Finally, the model only needs to predict one orientation value for each column of the input image to sufficiently represent all pedestrian pixels in the corresponding column.

The ColAtt-Net simultaneously divides the two tasks into object category segmentation and pedestrian orientation prediction. The model is trained in an end-to-end manner. As the pedestrian pixels in each vertical axis are sufficiently represented by only one orientation value, the ColAtt-Net puts a branch to predict the pedestrian orientations in column-wise working parallel with the branch that segments the object categories in pixel-level.

The conceptual approach proposed in this study is applicable to any base model. In this study, the ColAtt-Net uses PSPNet [20] as the base model. An extended version of PSPNet [22] was previously introduced for the attribute-aware semantic segmentation task and trained with the CityWalks dataset [43] which contains 23 classes consisting of 19 object categories (*road*, *building*, *car*, *person*, etc.) and four object's attributes (pedestrian orientations, i.e. *back*, *right*, *front*, and *left*). Here, the ColAtt-Net is proposed as a modification from the PSPNet basis, having a multi-tasking capability that improves the extended PSPNet in running the attribute-aware semantic segmentation task.

Figure 4 depicts the network architecture of the proposed ColAtt-Net model. The green blocks represent the same modules used in the base PSPNet model. It begins with some layers copied from the ResNet-101 [45], followed by a Pyramid Pooling Module

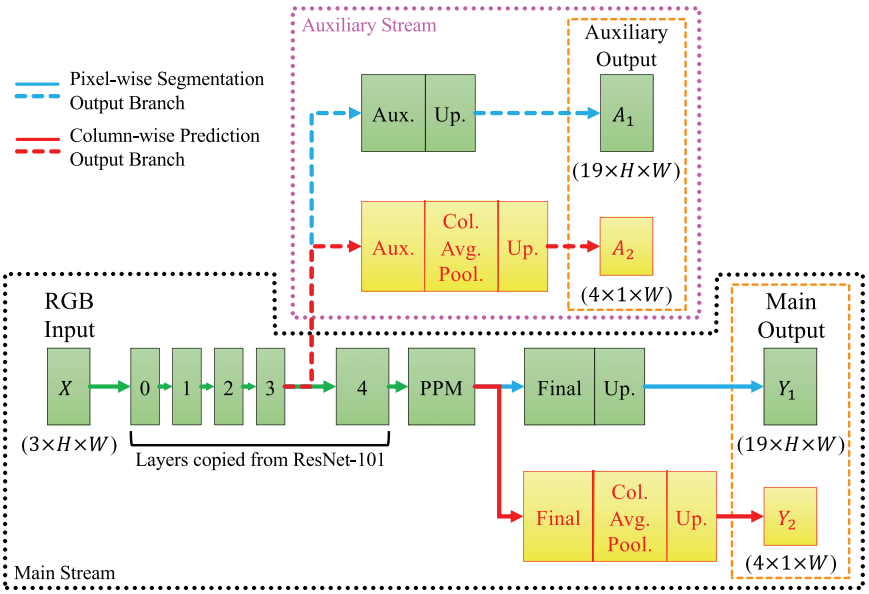


Fig. 4. Network architecture of the proposed ColAtt-net

(PPM), and enclosed with the segmentation output branch to yield the main output $Y_1$. In addition, the model has a 'side' stream or a so-called auxiliary stream that produces an auxiliary output. This is basically adopted from the original PSPNet model [20]. The auxiliary branch is to check the model's performance using the intermediary features given by the neural networks. Calculating loss from the auxiliary output helps improve the extracted image features in the middle of the neural networks, thus, accelerating the learning process. This mechanism will not interfere too much with the learning on the main stream since the auxiliary loss will be back-propagated to only a few front layers of the model. We extended the auxiliary branch in the proposed ColAtt-Net to produce two auxiliary outputs; $A_1$ represents the auxiliary output for the pixel-wise object segmentation task, while $A_2$ is for the column-wise orientation prediction task. By relying only on the main loss, a standard performance might be achievable. However, applying the auxiliary loss with an appropriate weight could further increase the performance of the trained model [20].

The yellow blocks show our main contributions with the proposed ColAtt-Net. It separates the task of pedestrian orientation prediction from the object category segmentation task by adding the output branches for the column-wise prediction of pedestrian orientations. These produce $Y_2$ at the main output and $A_2$ at the auxiliary output. In the segmentation branches that yield $Y_1$ and $A_1$, the size of the output channel is set to 19; Each result in a $19 \times H \times W$ tensor. Meanwhile, the column-wise prediction branch at the main stream adopts the final and up-sample modules, inserts a column average pooling module to shrink the size in the vertical axis but still keeping the spatial information of the tensor, and then adjusts the up-sample module size to yield $Y_2$ with the shape of $4 \times 1 \times W$. Its channel's size, i.e. 4, corresponds to the number of pedestrian orientations. The same modification is applied to the column-wise prediction branch at the auxiliary stream that produces $A_2$. A column average pooling module is inserted between the adapted auxiliary and up-sample modules. Note that these auxiliary outputs ($A_1$ and $A_2$) are utilized together with the main output only in the training stage. While in the testing stage, the ColAtt-Net uses only the main output ($Y_1$ and $Y_2$).

### 3.3. Converting the ground truth for training stage

The CityWalks dataset [43] provides sets of ground truths that contain 23 classes, including 19 categories of objects (labeled from 0 to 18) and 4 attributes of pedestrian orientations (labeled from 19 to 22). Here, the original ground truth ($G$) needs to be converted separately into a 19-class ground truth for the object segmentation ($G^{'}$) and a 4-class ground truth for the column-wise orientation prediction ($G^{''}$). $G$ is a two-dimensional matrix with the size of $H \times W$ and contains integers ranging from 0 to 22. As the conversion result, $G^{'}$ has the same size as $G$ but with labels ranging from 0 to 18 that correspond to the object categories, while $G^{''}$ is a matrix with the size of $1 \times W$ and labels ranging from 0 to 3 that correspond to the pedestrian orientations.

First, to obtain $G^{'}$, it is simply performed by overwriting all orientation labels in $G$ (19–22) with the label for the class *person*, i.e. 11. Second, to convert from $G$ to $G^{''}$, we need to find one orientation label that represents the target for each column in $G$. The chosen orientation is the most frequent label of orientation which is observed in a window of pixels. If two or more labels become the most frequent orientation, then one of them is selected randomly. In the implementation, we propose two choices of method, 'option 1' and 'option 2'. With option 1, the orientation target is found in every single-column window, whereas with option 2, the orientation target is found in every three-column window. Figure 5 illustrates this mechanism in more detail.

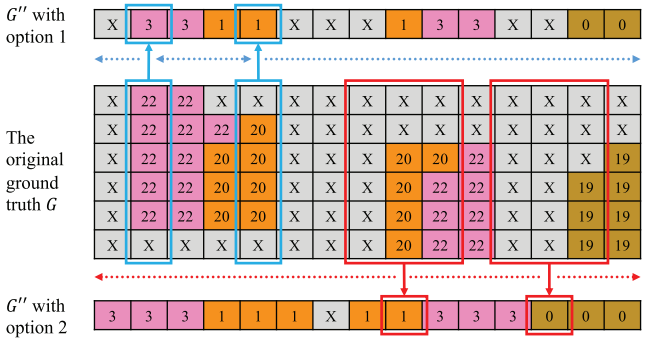


Fig. 5. Finding the orientations target for ground-truth $G^{''}$ with two options; all orientation labels are converted from [19, 22] to [0.3], respectively; this is executed along the column (notated with the dotted lines)

The process of finding the most frequent orientation label is only for pedestrian orientations classes. If there is not a single orientation label in the window, then the corresponding column for $G^{''}$ is assigned with the ignored label (255) and hence will not be involved in the loss calculation. From these converted ground truths, $G^{'}$ is used to train the ColAtt-Net's branch for the object category segmentation task, while $G^{''}$ is for the column-wise orientation prediction task.

### 3.4. Calculating the loss

In the training loop stage, the ColAtt-Net considers both outputs: main and auxiliary; each yields two sub-outputs, which are the 19-class object segmentation and the 4-class column-wise attribute prediction. From each branch, we calculate the respective loss using a standard cross-entropy loss function [46]; $\mathcal{L}_1$ is for the object segmentation loss and $\mathcal{L}_2$ is for the column-wise orientation prediction. Weights of the loss, $\beta_1$ and $\beta_2$, are applied to multiply each loss, respectively. These weights are important parameters in the training process to determine how much the portion of each loss will be counted for the back-propagated loss.

$$\mathcal{L}^x = \beta_1 \mathcal{L}_1^x + \beta_2 \mathcal{L}_2^x, \quad x \in \{\text{main}, \text{aux}\} \tag{1}$$

After calculating losses for both main and auxiliary outputs, we combine them to obtain the final loss $\mathcal{L}$ as formulated in Equation 2 and use it for the backward propagation; A constant of $\alpha$ is applied for the auxiliary loss to balance these two losses. Figure 6 shows the function that elaborates how this loss calculation works.

$$\mathcal{L} = \mathcal{L}^{\text{main}} + \alpha \mathcal{L}^{\text{aux}} \tag{2}$$

### 3.5. Post-processing for the final output

In the testing stage, the ColAtt-Net only considers the main output consisting of $Y_1$ and $Y_2$, and omits the auxiliary outputs. As previously explained, $Y_1$ represents the semantic segmentation output that contains 19 classes of object categories of which one of them is *person*. Meanwhile, $Y_2$ predicts the pedestrian's orientations in

**function** colatt_loss($Y_1$, $Y_2$, $A_1$, $A_2$, $G'$, $G''$)

$\mathcal{L}_1^{\text{main}}$ = cross_entropy_loss($Y_1$, $G'$)

$\mathcal{L}_1^{\text{aux}}$ = cross_entropy_loss($A_1$, $G'$)

$\mathcal{L}_2^{\text{main}}$ = cross_entropy_loss($Y_2$, $G''$)

$\mathcal{L}_2^{\text{aux}}$ = cross_entropy_loss($A_2$, $G''$)

$\mathcal{L}^{\text{main}} = \beta_1 \mathcal{L}_1^{\text{main}} + \beta_2 \mathcal{L}_2^{\text{main}}$

$\mathcal{L}^{\text{aux}} = \beta_1 \mathcal{L}_1^{\text{aux}} + \beta_2 \mathcal{L}_2^{\text{aux}}$

$\mathcal{L} = \mathcal{L}^{\text{main}} + \alpha \mathcal{L}^{\text{aux}}$

**return** $\mathcal{L}$, $\mathcal{L}^{\text{main}}$, $\mathcal{L}^{\text{aux}}$

Fig. 6. ColAtt-Net's loss calculation; cross entropy [46] is a common loss function for neural network training

column-wise. These two outputs are then combined in the post-processing step to yield the final output of the attribute-aware semantic segmentation. This is realized by overwriting all pixels classified as class *person* in $Y_1$ with the orientation class from the corresponding column in $Y_2$. Other classes are not overwritten with any orientation label. Figure 7 illustrates the post-processing mechanism to obtain the final output $Y$ from the proposed ColAtt-Net.

## 4. Experiment and Analysis

**4.1. Experimental setup** All experiments in this study used the CityWalks dataset [43] which consists of training and validation sets. Each contains 2975 and 500 finely-annotated images, respectively. The original size of the image is $1024 \times 2048$ pixels, but is resized in half per axis to $512 \times 1024$ pixels for processing speed reason.

For the training stage, we set a maximum iteration ($I_{\max}$) of 50 000 and a batch size ($b$) of 4. With the total number of training data ($N_{\text{train}}$) of 2975, every training routine will approximately run for a maximum epoch ($E_{\max}$) of 68. This number of epochs follows the equation below:

$$E_{\max} = \left\lceil \frac{I_{\max} \times b}{N_{\text{train}}} \right\rceil. \tag{3}$$

In addition, we set $\alpha$ to 0.4 because it empirically yielded optimal performance as described in the PSPNet paper [20].

To compare with the previous work, other training parameters such as initial learning rate, learning rate decay, weight decay, and momentum are set to 0.05, 0.9, 0.0001, and 0.9, respectively, which are used by the baseline method [22]. For the purposes of evaluation and observation, we analyze optimal parameters related to the ColAtt-Net, including the training option (1 or 2) to be used as well as the values assigned for $\beta_1$ and $\beta_2$.

To measure performance of each trained model on the validation set, we calculate the Intersection over Union (IoU) [21], especially for each orientation class, and the mean IoU for 4 orientations ($\text{mIoU}_4$), for 19 objects ($\text{mIoU}_{19}$), and for 22 object plus orientation classes ($\text{mIoU}_{22}$). To obtain $\text{mIoU}_{22}$, class *person* is substituted with four orientation classes. All scores are presented in percentage (%).

**4.2. Experimental result** First, we run some experiments to see which training options, between 1 and 2, is preferable to train the ColAtt-Net model. $\beta_1$ and $\beta_2$ are set with default values, which are 0.9 and 0.1, respectively. In addition, we run for each option with batch size $b$ of either 4 or 8. Performances of the trained model is measured with IoU and mIoU as previously explained. Table II shows the performance comparison of the trained models; bold numbers are the top scores over each column. Every training setting is run once with the default $I_{\max}$. From this experiment, we can see that option 2 generally performs better than option 1 and thus, option 2 is selected as the default training option for further experiments.

Next, an experiment is conducted to choose optimal values for the loss weights, with $\beta_1 \in \{0.8, 0.9, 1.0, 1.1, 1.2\}$ and $\beta_2 \in \{0.1, 0.2\}$. $\beta_2$ is set to a smaller value than $\beta_1$ because the trend shown in Fig. 8 indicates that $\mathcal{L}_2$ is always lower than $\mathcal{L}_1$. This is to maintain the proportional influences of these losses in training the two output branches according to their respective tasks. The performances are measured to observe how the values assigned to $\beta_1$ and $\beta_2$ influence both branches of the model's output, including the object semantic segmentation (represented in $\text{mIoU}_{19}$) and the column-wise orientation prediction (represented in $\text{mIoU}_4$ and the orientations' IoU). Table III shows these experimental results. We run one-time training for each pair of $\beta_1$ and $\beta_2$ values with default settings for $b$ and $I_{\max}$. Two scores for each column in the table are bold, showing the top-2 ranks for each performance metric. From the table, we can see that the pair of (0.9, 0.1) for ($\beta_1$, $\beta_2$) is the best for the column-wise orientation

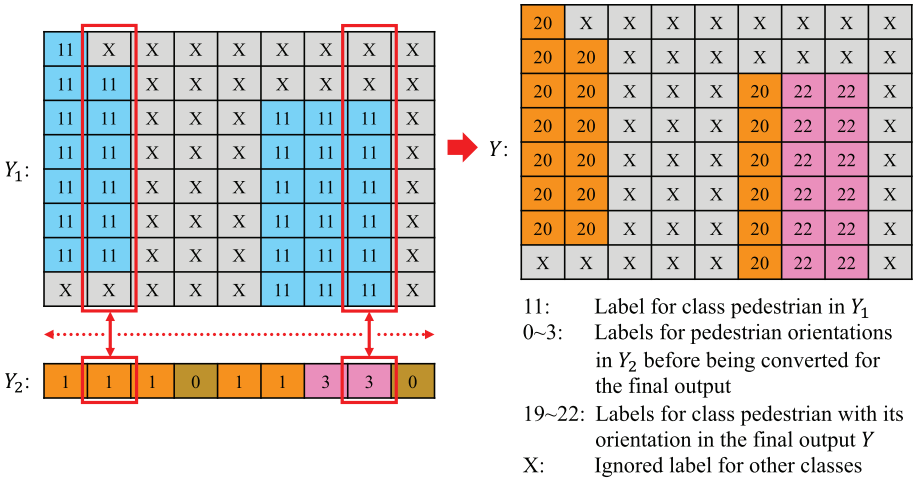


Fig. 7. Post-processing performed by the ColAtt-net

Table II. Observing the training options (1 or 2)

| Option | $B$ | $\text{mIoU}_4$ | $\text{IoU}_{\text{back}}$ | $\text{IoU}_{\text{right}}$ | $\text{IoU}_{\text{front}}$ | $\text{IoU}_{\text{left}}$ |
|---|---|---|---|---|---|---|
| 1 | 4 | 63.43 | 59.95 | **71.01** | 54.80 | 67.95 |
| 1 | 8 | 62.76 | 59.18 | 70.12 | 54.78 | 66.95 |
| 2 | 4 | **64.22** | 60.21 | 70.17 | **57.08** | **69.42** |
| 2 | 8 | 63.43 | **60.41** | 68.26 | 56.53 | 68.52 |

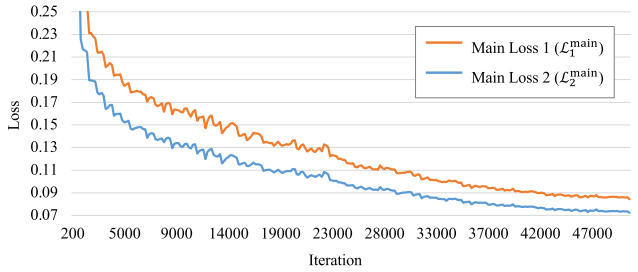


Fig. 8. The main losses for the two tasks during the training process; $\mathcal{L}_2^{\text{main}}$ is always below $\mathcal{L}_1^{\text{main}}$

prediction task, while to maintain the object semantic segmentation performance, (1.1, 0.1) deserves the best pair values for ($\beta_1$, $\beta_2$).

We then make comparisons between the method proposed in this study and the baseline method from the previous work as presented in Tables IV, V, and VI. For the proposed ColAtt-Net, three combinations for parameters $\beta_1$ and $\beta_2$ are chosen based on their achieved $\text{mIoU}_4$ and $\text{mIoU}_{19}$ in Table III; (0.9, 0.1) and (1.0, 0.1) become $\text{mIoU}_4$'s first and second ranks, respectively, while (1.1, 0.1) and (1.0, 0.1) achieve $\text{mIoU}_{19}$'s first and second ranks, respectively. Especially for the baseline method, the parameters $\beta_1$ and $\beta_2$ in Tables IV, V, and VI actually correspond to weight parameters of $\beta_o$ and $\beta_a$, respectively, and set with some pairs of values assigned to the baseline method in the previous work [22]. Each of the methods with its setting is trained 10 times to analyze the consistency of its performance.

From each method and its corresponding setting, we receive 10 performance results. Among those, first, we calculate the standard deviations as shown in Table IV. On the table's header, $\sigma_{22}$, $\sigma_{19}$, $\sigma_4$, $\sigma_{\text{back}}$, $\sigma_{\text{right}}$, $\sigma_{\text{front}}$, and $\sigma_{\text{left}}$ represent the standard deviations for $\text{mIoU}_{22}$, $\text{mIoU}_{19}$, $\text{mIoU}_4$, $\text{mIoU}_{\text{back}}$, $\text{mIoU}_{\text{right}}$, $\text{mIoU}_{\text{front}}$, and $\text{mIoU}_{\text{left}}$, respectively. Then, we calculate the averaged performances and show the results in Table V. Finally, we pick one out of 10 results, which represents the best record based on the achieved $\text{mIoU}_4$. For instance, we train ColAtt-Net with $\beta_1 = 0.9$ and $\beta_2 = 0.1$ 10 times, evaluate each of the trained models to obtain 10 records of performances, and then select the best one based on $\text{mIoU}_4$. All of the best records collected from each method and setting are shown in Table VI.

From Table IV, the proposed method generally produces a lower standard deviation than the baseline method. In Tables V and VI, one score is highlighted (bold) in each column to show which method gains the first rank for the corresponding performance metric. We can see that the ColAtt-Net's performance is slightly below the baseline's in segmenting the object categories as shown in $\text{mIoU}_{22}$ and $\text{mIoU}_{19}$ metrics. Other than that, the ColAtt-Net manages to excel far enough from the baseline method in terms of the $\text{mIoU}_4$ and the IoU of every orientation class. We investigated some validation results in a qualitative view and found several common cases that confirm the advantages of ColAtt-Net model over the baseline method, which will be discussed later in the next section.

**4.3. Discussions** In the first experiment, we observe the use of training option that represents how to find the orientation target for the ground truth, as shown in Fig. 5. The focus of this observation is on the model's branch for the column-wise orientation prediction which produces $Y_2$ and $A_2$. Table II indicates that applying option 2 generally yields better results compared to option 1, in terms of giving an accurate prediction of the pedestrian orientations. It is shown by $\text{mIoU}_4$ and the IoU of orientation classes.

From this experiment, we can realize that when the model is in the training process, especially to predict the orientation of a pedestrian in column-level, option 1 acts stricter as it focuses to find the target from exactly one column. Meanwhile, option 2 gives more relaxation as it also considers its neighboring columns. It means that option 2 has more spatial awareness, thus, it performs better and is preferable compared to option 1.

Table III. Observing the values assigned for $\beta_1$ and $\beta_2$

| $\beta_1$ | $\beta_2$ | Option | $B$ | $\text{mIoU}_{19}$ | $\text{mIoU}_4$ | $\text{IoU}_{\text{back}}$ | $\text{IoU}_{\text{right}}$ | $\text{IoU}_{\text{front}}$ | $\text{IoU}_{\text{left}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.8 | 0.1 | 2 | 4 | 69.77 | 61.72 | 58.63 | 68.17 | 52.90 | 67.18 |
| 0.8 | 0.2 | 2 | 4 | 67.28 | 61.59 | 57.26 | 65.36 | 54.59 | **69.16** |
| **0.9** | **0.1** | 2 | 4 | 69.91 | **64.22** | **60.21** | **70.17** | **57.08** | **69.42** |
| 0.9 | 0.2 | 2 | 4 | 68.23 | 61.01 | 56.73 | 66.51 | 53.64 | 67.15 |
| **1.0** | **0.1** | 2 | 4 | **70.27** | **63.11** | 59.38 | **70.00** | 54.70 | 68.34 |
| 1.0 | 0.2 | 2 | 4 | 69.37 | 61.92 | 58.37 | 67.84 | 54.23 | 67.242 |
| **1.1** | **0.1** | 2 | 4 | **71.22** | 62.57 | **60.35** | 67.69 | **55.75** | 66.50 |
| 1.1 | 0.2 | 2 | 4 | 70.19 | 61.49 | 57.83 | 68.55 | 54.37 | 65.21 |
| 1.2 | 0.1 | 2 | 4 | 70.09 | 61.84 | 59.33 | 67.94 | 54.20 | 65.89 |
| 1.2 | 0.2 | 2 | 4 | 69.80 | 60.11 | 58.29 | 63.63 | 52.95 | 65.58 |

Table IV. Standard deviations in each performance metric from multiple training conducted for the ColAtt-net and the baseline methods; this is particularly to confirm whether the performance of each method in several tries is stable or not

| Method | $\beta_1$ | $\beta_2$ | $\sigma_{22}$ | $\sigma_{19}$ | $\sigma_4$ | $\sigma_{\text{back}}$ | $\sigma_{\text{right}}$ | $\sigma_{\text{front}}$ | $\sigma_{\text{left}}$ |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 0.1 | 0.9 | 0.56 | 0.64 | 0.76 | 1.17 | 1.85 | 0.68 | 1.61 |
| Baseline | 0.5 | 0.5 | 0.40 | 0.43 | 0.95 | 1.71 | 1.82 | 0.73 | 1.21 |
| Baseline | 0.9 | 0.1 | 0.69 | 0.37 | 3.60 | 1.87 | 8.80 | 2.43 | 7.86 |
| ColAtt-Net | 0.9 | 0.1 | 0.60 | 0.68 | 0.97 | 0.91 | 1.09 | 0.00 | 1.51 |
| ColAtt-Net | 1.0 | 0.1 | 0.72 | 0.82 | 0.96 | 1.43 | 1.98 | 1.41 | 1.21 |
| ColAtt-Net | 1.1 | 0.1 | 0.43 | 0.50 | 1.00 | 1.41 | 1.37 | 1.32 | 1.83 |

Table V. Comparison of the averaged performances between the ColAtt-net and the baseline methods

| Method | $\beta_1$ | $\beta_2$ | $\overline{\text{mIoU}_{22}}$ | $\overline{\text{mIoU}_{19}}$ | $\overline{\text{mIoU}_4}$ | $\overline{\text{IoU}_{\text{back}}}$ | $\overline{\text{IoU}_{\text{right}}}$ | $\overline{\text{IoU}_{\text{front}}}$ | $\overline{\text{IoU}_{\text{left}}}$ |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 0.1 | 0.9 | **67.56** | 70.69 | 59.94 | 56.83 | 66.57 | 51.19 | 65.17 |
| Baseline | 0.5 | 0.5 | 67.50 | **70.89** | 59.33 | 55.29 | 66.59 | 51.04 | 64.39 |
| Baseline | 0.9 | 0.1 | 65.17 | 70.85 | 44.47 | 46.71 | 46.19 | 43.74 | 41.23 |
| ColAtt-Net | 0.9 | 0.1 | 67.03 | 69.77 | **62.18** | 59.04 | 67.97 | **54.34** | **67.34** |
| ColAtt-Net | 1.0 | 0.1 | 67.16 | 70.01 | 61.71 | 58.47 | **68.07** | 53.97 | 66.34 |
| ColAtt-Net | 1.1 | 0.1 | 67.42 | 70.23 | 62.03 | **59.28** | 67.85 | 53.83 | 67.16 |

In the next experiment, the results in Table III show that there is a trade-off between $\beta_1$ and $\beta_2$ in affecting the performances of two output branches. For $\beta_1$, there are conditions when $\text{mIoU}_{19}$ is increased but $\text{mIoU}_4$ is decreased, and *vise versa*. Whereas for $\beta_2$, we can see that assigning it with 0.1 performs better for both $\text{mIoU}_{19}$ and $\text{mIoU}_4$ compared to 0.2. However, we can still observe that the pair of (1.0, 0.1) is considered optimal for ($\beta_1$, $\beta_2$) based on $\text{mIoU}_{19}$ and $\text{mIoU}_4$ metrics in this experiment.

We also make comparisons between the proposed and the baseline methods. First, Table IV indicates the stable performances of the ColAtt-Net; All the standard deviation values are less than 2%. It is in contrast to the baseline method which may produce a relatively higher standard deviation. This indicates that the proposed method is more stable than the baseline method.

Next, Tables V and VI show the performance comparisons achieved by the proposed and the baseline methods. The concentration of the proposed method is divided into two sub-tasks as previously mentioned in Section 3. The ColAtt-Net's performance is slightly below the baseline method in the object segmentation task, indicated by $\text{mIoU}_{22}$ and $\text{mIoU}_{19}$ which both involve accuracies for the 19 object categories. In contrast, the ColAtt-Net outperforms the baseline method in the orientation prediction task thanks to the additional branch for predicting the pedestrian orientations in column-wise. The proposed method assigns an orientation label to every pixel of *person*, whether the pixel is correctly or incorrectly labeled as so. Nevertheless, the small differences in $\text{mIoU}_{22}$ and $\text{mIoU}_{19}$ indicate that the ColAtt-Net struggles to maintain its performance in the object segmentation task while increasing its accuracy in the orientations prediction task as indicated by $\text{mIoU}_4$ and each orientation's IoU. This generally shows that the proposed ColAtt-Net is able to improve the performance in predicting the pedestrian orientations, which means eliminating the ambiguity, without affecting the object category segmentation task too much.

Furthermore, qualitative results are investigated to confirm the advantages of the ColAtt-Net over the baseline method in some typical cases, shown in Fig. 9. In the figure, columns 1 and 2 show the case of single pedestrian facing right or left, where the baseline method outputs ambiguous orientations, while the ColAtt-Net successfully segments all the pedestrian pixels with a correct orientation. Column 3 exemplifies a common case where pedestrians cross the road in front of the in-vehicle camera viewpoint. In this condition, there is ambiguity in the baseline's output as we can see the legs of pedestrians classified in the wrong orientation. On the other hand, the proposed ColAtt-Net can handle these conditions well to correctly segment between pedestrians crossing to the left and to the right. The third typical case represented in columns 4 and 5 is a group of pedestrians walking on a sidewalk in the same direction. However, the baseline method failed to label all pedestrians correctly. This might be due to different looks and leg positions. Meanwhile, the ColAtt-Net had no ambiguity and managed to label all pedestrians in the correct direction thanks to its ability in examining the neighboring pedestrian pixels.

Table VI. Comparison of the top performances, based on the achieved $\text{mIoU}_4$, between the ColAtt-net and the baseline methods

| Method | $\beta_1$ | $\beta_2$ | $\text{mIoU}_{22}$ | $\text{mIoU}_{19}$ | $\text{mIoU}_4$ | $\text{IoU}_{\text{back}}$ | $\text{IoU}_{\text{right}}$ | $\text{IoU}_{\text{front}}$ | $\text{IoU}_{\text{left}}$ |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 0.1 | 0.9 | 66.90 | 69.84 | 61.01 | 58.21 | 67.63 | 51.08 | 67.11 |
| Baseline | 0.5 | 0.5 | **67.93** | **71.21** | 60.77 | 57.64 | 67.92 | 51.89 | 65.65 |
| Baseline | 0.9 | 0.1 | 65.55 | 70.43 | 49.19 | 46.77 | 55.45 | 45.13 | 49.39 |
| ColAtt-Net | 0.9 | 0.1 | 67.48 | 69.91 | **64.22** | 60.21 | 70.17 | **57.08** | 69.42 |
| ColAtt-Net | 1.0 | 0.1 | 67.59 | 70.27 | 63.11 | 59.38 | 70.00 | 54.70 | 68.34 |
| ColAtt-Net | 1.1 | 0.1 | 67.61 | 70.13 | 64.10 | **61.27** | **70.18** | 55.34 | **69.62** |

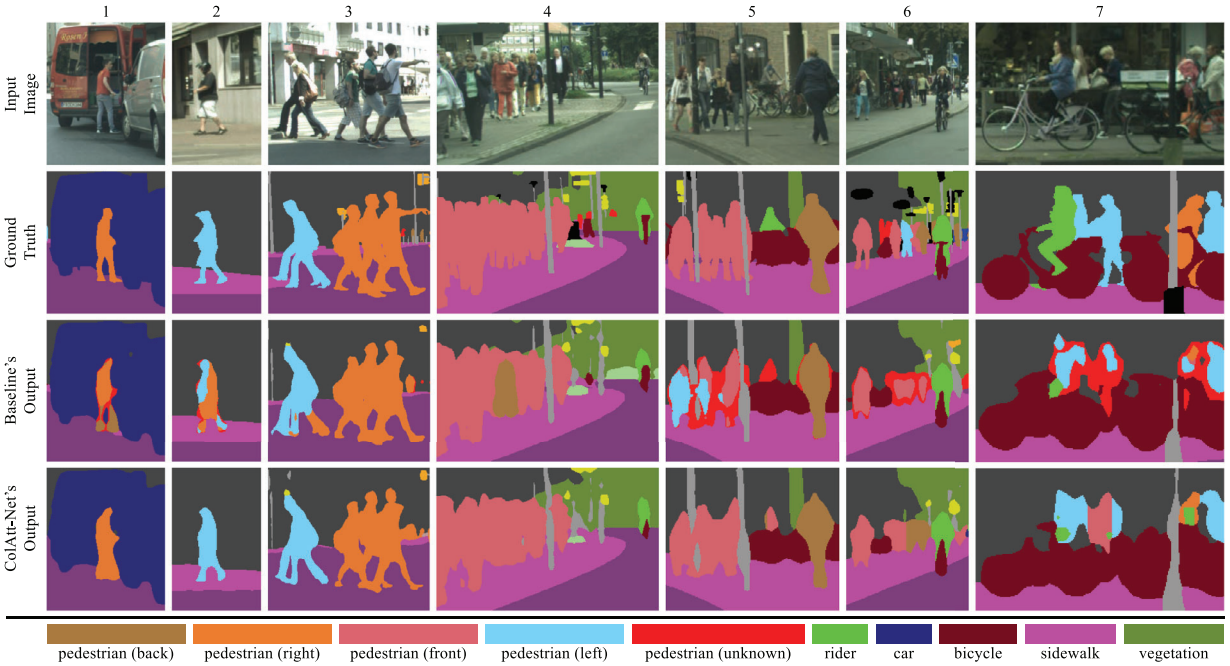


Fig. 9. Qualitative results and comparisons between the baseline and the proposed methods

Table VII. Observing the values assigned for the initial learning rate (μ) in training the ColAtt-net model

| $\mu$ | $\text{mIoU}_{22}$ | $\text{mIoU}_{19}$ | $\text{mIoU}_{4}$ | $\text{IoU}_{\text{back}}$ | $\text{IoU}_{\text{right}}$ | $\text{IoU}_{\text{front}}$ | $\text{IoU}_{\text{left}}$ |
|---|---|---|---|---|---|---|---|
| 0.025 | 65.01 | 67.62 | 61.96 | 56.71 | 68.55 | 54.01 | 68.58 |
| 0.01 | **68.25** | **71.02** | 62.86 | 59.35 | 70.09 | 54.27 | 67.72 |
| 0.005 | 67.48 | 69.91 | **64.22** | **60.21** | **70.17** | **57.08** | **69.42** |
| 0.0025 | 67.13 | 70.26 | 60.91 | 55.83 | 66.29 | 54.00 | 67.51 |
| 0.001 | 65.68 | 68.80 | 58.84 | 54.71 | 66.26 | 50.77 | 63.63 |

There is a case in column 6 where a group of pedestrians with three different orientations are not properly segmented by both methods. It is still difficult for the proposed method to handle crowds of pedestrians facing various orientations. Nevertheless, the ColAtt-Net is at least slightly better at separating pedestrians' orientations in object-level.

All of the abovementioned examples show that, in fact, the two methods are good at segmenting pedestrian objects, but the ColAtt-Net outperforms the baseline method in segmenting the pedestrian orientations. In addition, column 7 shows a case where both methods incorrectly segmented a rider into a pedestrian. In a condition that the cyclist is mixed with pedestrians, the two methods are still confused.

Moreover, we try to enhance the ColAtt-Net by optimizing one of its training parameters, which is the initial learning rate ($\mu$). We adjust $\mu$ to greater and smaller values, including 0.025, 0.01, 0.0025, and 0.001. For each of these values, a training process is executed once. We take the top result from the default $\mu$ of 0.005 and $(\beta_1, \beta_2) = (0.9, 0.1)$ as shown in Table VI for comparison. From Table VII, we can see that $\mu$ of 0.01 improves the ColAtt-Net in terms of $\text{mIoU}_{22}$ and $\text{mIoU}_{19}$, but still cannot surpass the default $\mu$ in terms of $\text{mIoU}_4$ and all orientations' IoU. From this, we could not find a value for $\mu$ that increases the performance of the ColAtt-Net in general. The $\mu$ of 0.005 is still considered a reasonable setting, particularly for the orientation classes.

The overall results in the experiments and observations show that the ColAtt-Net proposed in this study can improve the performance of the attribute-aware semantic segmentation tasks. It means that the proposed method successfully reduces the inaccuracy in segmenting the pedestrian orientations while still maintaining the accuracy in segmenting the object categories. By applying the ColAtt-Net to the applications related to an autonomous driving system, the reluctance to recognize pedestrian attributes can be drastically diminished. This is certainly important for a computer vision technique embedded in such a system so that it can accurately understand the traffic situation, especially the pedestrian movement, and thereby improve anticipation and prevention of the potential risks.

## 5. Conclusion

The limitations that occurred in the attribute-aware semantic segmentation, especially in traffic conditions, have been conveyed in this paper. Related to this, a novel model called ColAtt-Net was proposed to overcome the ambiguity in segmenting pedestrian orientations with the input images captured by a car-mounted camera. The ColAtt-Net is a deep neural network model based on the multi-task learning framework that simultaneously performs the two tasks including the object categories segmentation and the column-wise orientations prediction tasks. We also provided several

experiments to observe the optimal parameters for training the ColAtt-Net. It was demonstrated that the proposed method is effective to reduce the ambiguity in segmenting the pedestrian orientations in object-level, which is very important when applying to an autonomous driving system. For future work, optimizing ColAtt-Net's training parameters more comprehensively can also be considered for a possible improvement to the method's performance.

## Acknowledgment

The first author would like to express his gratitude to Indonesia's BUDI-LN scholarship for the generous supports to his PhD study in Nagoya University, Japan. Parts of this research were also supported by MEXT, KAKENHI Grant Number JP 17H00745.

## References

(1) Kemker R, Salvaggio C, Kanan C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing* 2018; **145**:60–77.

(2) Kayalibay B, Jensen G, van der Smagt P. CNN-based segmentation of medical imaging data. *Computing Research Repository*, arXiv:1701.03056 2017. https://arxiv.org/abs/1701.03056.

(3) Jiang F, Grigorev A, Rho S, Tian Z, Fu Y, Jifara W, Adil K, Liu S. Medical image semantic segmentation based on deep learning. *Neural Computing and Applications* 2018; **29**(5):1257–1265.

(4) Xia KJ, Yin HS, Zhang YD. Deep semantic segmentation of kidney and space-occupying lesion area based on SCNN and ResNet models combined with SIFT-flow algorithm. *Journal of Medical Systems* 2019; **43**(1):2.

(5) Kim W, Seok J. Indoor semantic segmentation for robot navigating on mobile. *Proceedings of 2018 International Conference on Ubiquitous and Future Networks*. pp. 22–25 (2018)

(6) Milioto A, Lottes P, C. Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs, *Proceedings of 2018 IEEE International Conference on Robotics and Automation*. pp. 2229–2235 (2018)

(7) Zhang Y, Chen H, He Y, Ye M, Cai X, Zhang D. Road segmentation for all-day outdoor robot navigation. *Neurocomputing* 2018; **314**:316–325.

(8) Treml M, Arjona-Medina J, Unterthiner T, Durgesh R, Friedmann F, Schuberth P, Mayr A, Heusel M, Hofmarcher M, Widrich M, Nessler B. Speeding up semantic segmentation for autonomous driving, *Proceedings of Machine Learning for Intelligent Transportation Systems, Neural Information Processing Systems Workshop*. **2**, p.7 (2016)

(9) Luc P, Neverova N, Couprie C, Verbeek J, LeCun Y: Predicting deeper into the future of semantic segmentation, *Proceedings of 2017 IEEE International Conference on Computer Vision*. pp. 648–657 (2017)

(10) Li L, Qian B, Lian J, Zheng W, Zhou Y. Traffic scene segmentation based on RGB-D image and deep learning. *IEEE Transactions on Intelligent Transportation Systems* 2017; **19**(5):1664–1669.

(11) Li B, Liu S, Xu W, Qiu W. Real-time object detection and semantic segmentation for autonomous driving. *Proceedings of SPIE, MIPPR 2017: Automatic Target Recognition and Navigation* 2018; **10608**:167–174.

(12) Tseng Y-H, Jan S-S. Combination of computer vision detection and segmentation for autonomous driving. *Proceedings of 2018 IEEE/ION Position, Location and Navigation Symposium*. pp. 1047–1052 (2018)

(13) Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *Proceedings of 2015 IEEE Conerence. on Computer Vision and Pattern Recognition*. pp. 3431–3440 (2015)

(14) Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2017; **39**(12):2481–2495.

(15) Kendall A, Badrinarayanan V, Cipolla R. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *Computing Research Repository*, arXiv:1511.02680 2015. https://arxiv.org/abs/1511.02680.

(16) Zhao H, Qi X, Shen X, Shi J, Jia J. ICNet for real-time semantic segmentation on high-resolution images. *Proceedings of 2018 European Conference on Computer Vision*, Part III. pp. 418–434 (2018)

(17) Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2018; **40**(4):834–848.

(18) Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. *Computing Research Repository*, *arXiv* 1706.05587 2017. https://arxiv.org/abs/1706.05587.

(19) Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of 2018 European Conerence on Computer Vision*. Part VII, pp. 833–851 (2018)

(20) Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2881–2890 (2017)

(21) Sulistiyo MD, Kawanishi Y, Deguchi D, Hirayama T, Ide I, Zheng JY, Murase H. Attribute-aware semantic segmentation of road scenes for understanding pedestrian orientations. *Proceedings of 21st IEEE Inernational Conference on Intelligent Transportation Systems*. pp. 2698–2703 (2018)

(22) Sulistiyo MD, Kawanishi Y, Deguchi D, Ide I, Hirayama T, Zheng JY, Murase H. Attribute-aware loss function for accurate semantic segmentation considering the pedestrian orientations. *IEICE Transactions Fundamentals* 2020; **E103-A**(1):231–242.

(23) Sulistiyo MD, Kawanishi Y, Deguchi D, Ide I, Hirayama T, Murase H. Performance boost of attribute-aware semantic segmentation via data augmentation for driver assistance. *Proceedings of 8th IEEE Conference on Information and Communication Technology*. 6p. (2020)

(24) Sulistiyo MD, Kawanishi Y, Deguchi D, Ide I, Murase H: A preliminary study of attribute-aware semantic segmentation for pedestrian understanding. *Proceedings of* 2017 *Electric/Electronic/Information Engineering Related Society Tokai Sectors Joint Convention*. no.A2-7, 1p. (2017)

(25) Kawanishi Y, Deguchi D, Ide I, Murase H, Fujiyoshi H: Misclassification tolerable learning for robust pedestrian orientation classification. *Proceedings of 23rd IAPR Inernational Conference on on Pattern Recognition*. pp. 486–491 (2016)

(26) Shinmura F, Kawanishi Y, Deguchi D, Hirayama T, Ide I, Murase H, Fujiyoshi H. Estimation of driverâ€™s insight for safe passing based on pedestrian attributes. *Proceedings of 21st IEEE Inernational Conference on Intelligent Transportation Systems*. pp. 1041–1046 (2018)

(27) Hariyono J, Jo KH. Detection of pedestrian crossing road: A study on pedestrian pose recognition. *Neurocomputing* 2017; **234**:144–153.

(28) Deng Y, Luo P, Loy CC, Tang X. Pedestrian attribute recognition at far distance. *Proceedings of 22nd ACM Inernational Conference on on Multimedia*. pp. 789–792 (2014)

(29) Li D, Chen X, Huang K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. *Proceedings of 3rd Asian Conference on Pattern Recognition*. pp. 111–115 (2015)

(30) Li D, Zhang Z, Chen X, Ling H, Huang K. A richly annotated dataset for pedestrian attribute recognition. *Computing Research Repository*, *arXiv*:1603.07054 2016. https://arxiv.org/abs/1603.07054.

(31) Schneemann F, Heinemann P. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments.

*2016 IEEE/RSJ Inernational Conference on Intelligent Robots and Systems*. pp. 2243–2248 (2016)
(32) Brostow GJ, Fauqueur J, Cipolla R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* 2009; **30**(2):88–97.
(33) Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research* 2013; **32**(11):1231–1237.
(34) Neuhold G, Ollmann T, Rota Bulò S, and Kontschieder P. The Mapillary vistas dataset for semantic understanding of street scenes. *Proceedings of 16th IEEE Inernational Conference on on Computer Vision*. pp. 4990–4999 (2017)
(35) Cordts M, Omran M, Ramos S, Scharwächter T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset. *Proceedings of CVPR* 2015 *Workshop on the Future of Datasets in Vision*. 4p. (2015)
(36) Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. pp3213–3223 (2016)
(37) Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation. *Computing Research Repository arXiv preprint*, arXiv:1704.06857 2017. https://arxiv.org/abs/1704.06857.
(38) Guo Y, Liu Y, Georgiou T, Lew MS. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval* 2018; **7**:87–93.
(39) Liu X, Deng Z, Yang Y. 2019: Recent progress in semantic image segmentation. *Artificial Intelligence Review* 2019; **52**(2):1089–1106.
(40) He K, Gkioxari G, Dollár P, R. Girshick. Mask R-CNN. *Proceedings of 2017 IEEE Inernational Conference on on Computer Vision*. pp. 2961–2969 (2017)
(41) Zhao J, Li J, Cheng Y, Sim T, Yan S, Feng J: Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing, *Proceedings of 26th ACM Inernational Conference on on Multimedia*, pp. 792–800 (2018)
(42) Kirillov A, He K, Girshick R, Rother C, Dollár P. Panoptic segmentation. *Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9404–9413 (2019)
(43) Sulistiyo MD, Kawanishi Y, Deguchi D, Ide I, Hirayama T., Murase H. CityWalks: An extended dataset for attribute-aware semantic segmentation. *Proceedings of* 2019 *Electric/Electronic/Information Engineering Related Society Tokai Sectors Joint Convention*. no. B1-1, 1p. (2019)
(44) Flohr F, Dumitru-Guzu M, Kooij JFP, Gavrila DM. A probabilistic framework for joint pedestrian head and body orientation estimation. *IEEE Transactions on Intelligent Transportation Systems* 2015; **16**(4):1872–1882.
(45) He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016)
(46) Pohlen T, Hermans A, Mathias M, Leibe B. Full-resolution residual networks for semantic segmentation in street scenes. *Proceedings of 2017 IEEE Conerence. on Computer Vision and Pattern Recognition*. pp. 4151–4160 (2017)

**Mahmud Dwi Sulistiyo** (Non-member) received his Bachelor (S.T.) and Master's (M.T.) degrees in Informatics Engineering from Telkom Institute of Technology, Indonesia, in 2010 and 2012, respectively. From 2010 to 2013, he became an Honorary Tutor at Telkom Institute of Technology. Since 2013, at the same time Telkom Institute of Technology changed its name to Telkom University, he has been appointed as a Permanent Lecturer at School of Computing. Starting from 2017, he has been studying his Ph.D in Department of Intelligent Systems at Graduate School of Informatics, Nagoya University, Japan. His current research focuses on attribute-aware semantic segmentation.

**Yasutomo Kawanishi** (Non-member) received his B.Eng degree in Engineering and M.Inf and Ph.D degrees in Informatics from Kyoto University, Japan, in 2006, 2008, and 2012, respectively. He became a Post Doctoral Fellow at Kyoto University, Japan in 2012. He moved to Nagoya University, Japan as a Designated Assistant Professor in 2014. In 2015, he became an Assistant Professor, and since 2020, he has been a Lecturer there. His main research interest is computer vision for human understanding, which includes pedestrian detection, tracking, retrieval, and recognition, for surveillance and in-vehicle videos. He received the best paper award from SPC2009, and Young Researcher Award from IEEE ITS Society Nagoya Chapter. He is a member of IEEE, IIEEJ, and IEICE.

**Daisuke Deguchi** (Non-member) received his B.Eng and M.Eng degrees in Engineering and Ph.D degree in Information Science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He became a Post Doctoral Fellow at Nagoya University, Japan in 2006. From 2008 to 2012, he had been an Assistant Professor at the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor in Information Strategy Office, and Since 2020, he has been an Associate Professor at the Graduate School of Informatics. He is working on the object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE, IPS Japan and IEEE.

**Ichiro Ide** (Non-member) received his B.Eng., M.Eng. and Ph.D from The University of Tokyo in 1994, 1996 and 2000, respectively. He became an Assistant Professor at the National Institute of Informatics, Japan in 2000. From 2004 to 2019, he was an Associate Professor at Nagoya University, Japan. Since 2020, he has been a Professor. He was also a Visiting Associate Professor at National Institute of Informatics from 2004 to 2010, an Invited Professor at Institut de Recherche en Informartique et Systèmes Aláeatoires (IRISA), France in 2005, 2006 and 2007, a Senior Visiting Researcher at ISLA, Instituut Voor Informatica, Universiteit van Amsterdam from 2010 to 2011. His research interest ranges from the analyses and indexing to re-targeting of multimedia contents, especially in large-scale broad-cast video archives, mostly on news, cooking, and sports contents. He is a senior member of IEICE and IPS Japan, and also a member of JSAI, IEEE, and ACM.

**Takatsugu Hirayama** (Non-member) received the M.E. and D.E. degrees in Engineering Science from Osaka University in 2002 and 2005, respectively. From 2005 to 2011, he was a Research Assistant Professor at the Graduate School of Informatics, Kyoto University. He is currently a Designated Associate Professor at the Institutes of Innovation for Future Society, Nagoya University. His research interests include computer vision (face recognition, gaze tracking, visual attention modeling, action recognition) and human-computer interaction (multi-modal interaction design, internal state estimation, interaction dynamics analyses). He has received the best paper award from IEICE ISS in 2014. He is a member of IEEE, ACM, IEICE, and IPS Japan.

**Hiroshi Murase** (Non-member) received his B.Eng, M.Eng, and Ph.D degrees in Electrical Engineering from Nagoya University, Japan. In 1980 he joined the Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993 he was a Visiting Research Scientist at Columbia University, New York. From 2003 he is a Professor of Nagoya University, Japan. He was awarded the IEICE Shinohara Award in 1986, the Telecom System Award in 1992, the IEEE CVPR (Conference on Computer Vision and Pattern Recognition) Best Paper Award in 1994, the IPS Japan Yamashita Award in 1995, the IEEE ICRA (International Conference on Robotics and Automation) Best Video Award in 1996, the Takayanagi Memorial Award in 2001, the IEICE Achievement Award in 2002, and the Ministry Award from the Ministry of Education, Culture, Sports, Science and Technology in 2003. He is a Fellow of IEEE, IEICE, and IPS Japan.